על ייחודיות פרופיל ה־DNA
תיאוריה ותוצאות מסימולציה ממוחשבת של פרופילים גנטיים באוכלוסיה רחבה

מרדכי הלפרט*

תקציר

מחקר זה ממחיש באמצעות חישובים תיאורטיים וסימולציות ממוחשבות את המשמעות של הסתברות התאמה מקרית בראיית ה־DNA ואת התנאים המתמטיים הרלוונטיים להכרזה על ייחודיות הפרופיל. המחקר עונה על שתי שאלות: (א) כמה זוגות בני אדם, שלישיות וכו' יש באוכלוסיה שגודלה כגודל אוכלוסיית מדינת ישראל אשר להם פרופיל גנטי זהה? (ב) האם ישנה אפשרות ממשית שאדם שהפרופיל הגנטי שלו מצוי במאגר נתונים גדול שבידי הרשויות, יופלל כתוצאה מפשע שביצע אדם שהפרופיל הגנטי שלו לא נמצא במאגר? הנחת המחקר היא כי מקור הפרופילים הגנטיים באוכלוסיה יהודית הומוגנית. המחקר אינו מביא בחשבון קיום של תת־אוכלוסיות, קרבת משפחה ושגיאות מעבדה, ובכך התוצאות מהוות רק חסם תחתון לתשובות על שאלות המחקר. כדי לענות על השאלה הראשונה נעשה שימוש במחקר בתוצאות תיאורטיות וכן בסימולציה ממוחשבת של פרופילים גנטיים אשר נבנו על ידי תוכנת מחשב בהתאם לשכיחות אללים באוכלוסיה היהודית בישראל. המחקר מוצא כי בקרב 7 מיליון פרופילים גנטיים המורכבים מתשעה אתרים ואשר מתפלגים על פי שכיחות אללים באוכלוסיה יהודית בישראל, יש בממוצע כ־14,865 זוגות פרופילים זהים. בתוכם, יש מאות שלישיות וכמה רביעיות בודדות של פרופילים זהים. כאשר מדובר על פרופילים גנטיים המורכבים משישה אתרים בלבד, יש מספר עצום של כפילויות ובפרט, יכול שפרופיל גנטי יהיה משותף למאות בני אדם. כדי לפתור את השאלה השנייה, המחקר משתמש בכלים תיאורטיים ובסימולציה ממוחשבת של חקירת פשעים רבים למול מאגר המכיל מיליון פרופילים גנטיים. התשובה לשאלה השנייה היא שקיימת אפשרות מוחשית להפללת אדם שהפרופיל הגנטי שלו מצוי במאגר נתונים גדול, כתוצאה מפשע שביצע אדם שהפרופיל הגנטי שלו לא נמצא במאגר, גם כאשר הסתברות ההתאמה המקרית אחת למיליארדים רבים וגם שמדובר בפרופילים המורכבים מתשעה אתרים. התוצאות נוספות העולות מהמחקר הינן: (א) עבור פרופילים גנטיים המורכבים משישה אתרים נצפה בממוצע כי כבר בקרב 1,140 פרופילים יימצא זוג פרופילים זהה. עבור פרופילים גנטיים המורכבים מתשעה אתרים נצפה בממוצע כי כבר בקרב 57,413 פרופילים יימצא זוג פרופילים זהה. (ב) כדי לקבוע ייחודיות של פרופיל גנטי באופן שבממוצע יהיו רק 0.01 זוגות פרופילים זהים בכל כדור הארץ המכיל כשבעה מיליארד בני אדם יש צורך בהסתברות התאמה מקרית ממוצעת של $4.08x10^{-22}$. (ג) כדי לקבוע ייחודיות של פרופיל גנטי באופן שבממוצע יהיו רק 0.01 זוגות פרופילים כאלה בישראל בלבד המונה כשבעה מיליון תושבים הארץ יש צורך בהסתברות התאמה מקרית ממוצעת של $4.08x10^{-16}$. מסקנות המחקר הינן כי פרופילים גנטיים המופקים מתשעה אתרים אינם מספיקים כדי לקבוע את ייחודיות הפרופיל הגנטי ולכן יש להגדיל את מספר האתרים בהם מתבצעת הבדיקה ובהתאמה, לבצע מחקר אשר יספק נתונים על שכיחות האללים באתרים הנוספים, באוכלוסיה. מסקנה נוספת כי על הרשויות למצוא את הדרך אשר מחד תשמור על פרטיות האנשים אשר הפרופיל הגנטי שלהם מצוי במאגר ומצד שני, תאפשר לחוקרים לבחון את התיאוריה הפורנזית למול הנתונים המצויים במאגר.

* ד"ר מרדכי הלפרט, Ph.D., פיסיקאי העוסק במחקר ופיתוח בתעשייה: mhalpert@zahav.net.il

א. הקדמה

בספרות העוסקת בראיית ה־DNA, מקובל לדון בשתי אפשרויות של מציאת חשוד בעבירה.[1] האחת מכונה מקרה האימות (confirmation case).[2] במקרה זה, ראיות מסוימות קושרות חשוד לאירוע פלילי ומטרת הבדיקה לאמת או לשלול את האפשרות שהחשוד הוא העבריין. אפשרות אחרת כאשר הפרופיל הגנטי של העבריין כפי שהופק מהדגימה שנלקחה מזירת הפשע, נסרק במאגר פרופילים גנטיים, ונמצא חשוד אשר לו פרופיל גנטי זהה. צורת הגעה כזו לחשוד מכונה לעיתים Cold Hit או "trawl case".[3] המקרה השני, עלול להתאפיין במקרים מסוימים בהיעדר ראיות מלבד בדיקת ה־DNA, כלפי אותו חשוד.

בבתי המשפט בישראל ובעולם, חוזק ראיית ה־DNA מבוטא בהסתברות ההתאמה המקרית (– Random Match Probability RMP). הסתברות זאת, היא ההסתברות כי בהינתן שאדם חף מפשע, תימצא התאמה בין פרופיל הנאשם לפרופיל העבריין.[4] בעניין מוראד אבו חמאד כתב השופט חשין כך:

> **"שכיחות של אחד בין מיליארדים די בה כדי לקשור פלוני לעבירה וכך אף בשכיחות של מיליונים, ואולם אין צורך שנקבע מסמרות."[5]**

לבדיקת טענה זאת כי שכיחות של מיליארדים ואף אולי של מיליונים מספקת כדי לקשור פלוני לעבירה, נבחנה במחקר שאלת ייחודיות הפרופיל הגנטי. בפרט, המחקר עונה על השאלה כמה זוגות, שלישיות ורביעיות וכו' של פרופילים גנטיים תואמים יש באוכלוסיה גדולה כגודל אוכלוסיית מדינת ישראל. הדבר נעשה הן בצורה תיאורטית והן בצורה של סימולציה נומרית של שבעה מיליון פרופילים גנטיים המורכבים משישה ותשעה אתרים, המתפלגים בהתאם לשכיחות אללים באוכלוסיה היהודית בישראל.[6] לאחר מכן, נערכה סימולציה ממוחשבת של חקירות פשעים רבים על ידי המשטרה אל מול מאגר DNA המונה מיליון פרופילים. תוצאות הסימולציה מראות כי במהלך סריקות רבות כאלה, ישנה אפשרות ריאלית שפרופיל גנטי של אדם הנמצא בתוך המאגר יהיה זהה לפרופיל גנטי של עבריין שלא נמצא במאגר.

ב. הנחות המחקר

המחקר עונה על שתי שאלות:

א. כמה זוגות בני אדם, שלישיות וכו' יש באוכלוסיה שגודלה כגודל אוכלוסיית מדינת ישראל אשר להם פרופיל גנטי זהה?

ב. האם ישנה אפשרות ממשית שאדם שהפרופיל הגנטי שלו מצוי במאגר נתונים גדול שבידי הרשויות, יופלל כתוצאה מפשע שביצע אדם שהפרופיל הגנטי שלו לא נמצא במאגר?

ההנחות עליהם מבוסס המחקר הנוכחי הן אלה:

1. הפרופילים הגנטיים שייכים לאוכלוסיה הומוגנית יהודית בלבד.
2. אין הבאה בחשבון של תת־אוכלוסיות וקרובי משפחה.

---

1 בדיקת ה־DNA מתוארת רבות ובפירוט רב בספרות. להעמקה ראו אחיקם סטולר ויורם פלוצקי "D.N.A. על דוכן העדים" **רפואה ומשפט** 25, 143 (2001). ראו גם נירה גלילי ואסא מרבך "אנליזה של דנ"א למטרות פורנזיות" **פלילים** ב 225 (1991). ראו גם את החלק העוסק בבדיקת DNA במדריך האמריקאי לראיות מדעיות: Fed Judicial Ctr., *Reference Manual on Scientific Evidence* 487–576 (2nd ed., 2000).

2 Peter Donnelly & Richard D Friedman, *DNA Database Searches and the Legal Consumption of Scientific Evidence*, 97 Mich. L. Rev. 931, 932 (1999).

3 שם.

4 Boaz Sangero and Mordechai Halpert, *Why a Conviction Should Not Be Based on a Single Piece of Evidence: A Proposal for Reform*, 48 Jurimetrics J. 43, 72 (2007) ראו גם מרדכי הלפרט ומשה פרדס, "האומנם ניתן להרשיע על בסיס ראיה מדעית יחידה? המקרה של ראיות טביעת האצבע ו־DNA", **עיוני משפט** ל 399 (2007);

5 ע"פ 9724/02 **מוראד אבו־חאמד נ' מדינת ישראל**, דינים עליון כרך ס"ה, 552 (2003) בפס' 35 לפסק־דינו של השופט חשין.

6 כידוע, בישראל חיים גם יהודים וגם ערבים. התפלגות האללים שונה מאוכלוסיה לאוכלוסיה. למרות זאת, בחרנו באוכלוסיה היהודית בלבד לצורך הפשטות בלבד.

3. אין הבאה בחשבון של שגיאות מעבדה.[7]

אכן, הנחות אלה אינן מתארות את המציאות. כך למשל, תיתכן גם אפשרות של קירבת משפחה ברמה של אח חורג אשר אינה ידועה כלל לסביבה, ייתכנו מקרים בהם אדם אינו מודע כלל לעובדה כי הוא אב (למשל, כתוצאה מקיום יחסי־מין מחוץ לנישואין) וכן ייתכנו מקרים בהם האם אינה יודעת בוודאות מי האב. עובדה זאת חשובה מאוד למשפט הפלילי. אולם הנחות מחקר אלה הן מינימאליות באופן שהתשובות לשאלות המחקר מהוות חסם תחתון לכמות ההתאמות שתימצא בפועל. היינו, אנו נצפה במציאות ליותר התאמות מאלה שתוצאות מחקר זה מצביע עליהם.

ג. כמה פרופילים גנטיים זהים יש בקרב אוכלוסיה המונה שבעה מיליון בני אדם?

ההנחה הבסיסית היא כי DNA אצל בני האדם, פרט לתאומים זהים, הוא ייחודי.[8] אולם פרופיל גנטי אינו כל ה־DNA, אלא ייצוג ממנו המתקבל ממספר אתרים במולקולת ה־DNA.[9] מכאן עולה שאלת המחקר, כמה פרופילים גנטיים זהים יש בקרב אוכלוסיה שגודלה כגודל מדינת ישראל? לצורך מחקר זה בלבד, בחרנו בפרופילים גנטיים המתפלגים על פי שכיחות אללים באוכלוסיה היהודית בישראל. תשובה מדויקת לשאלה הזאת יכולה להינתן רק אם יהיו בידינו הנתונים האמיתיים[10] על כלל הפרופילים הגנטיים באוכלוסיה. למרות זאת, ניתן להעריך הן בצורה תיאורטית והן בצורה של סימולציה נומרית, כמות זאת.

תחת הנחות המחקר, כמות ההתאמות הממוצעת התיאורטית $\bar{n}$, שנקבל בקרב אוכלוסיה המונה N אנשים המאופיינת בהסתברות התאמה מקרית ממוצעת $\bar{p}$ היא:[11]

$$(1)\ \bar{n} = \frac{N \text{ x } (N-1)}{2} \text{ x } \bar{p}$$

כלומר, כמות הזוגות התואמים הממוצעת באוכלוסיה בגודל N עולה בקירוב באופן ריבועי עם גידול האוכלוסיה. אציין כי בחישוב זה, שלישיית התאמות תחשוב כשלושה זוגות תואמים[12] ובאותו אופן, רביעית התאמות תחשב כשישה זוגות וכולי.[13,14]

כדי לשכנע את הקורא כי משוואה (1), למרות פשטותה, היא נכונה, וכמקרה מבחן, בחרתי להדגימה על ואריאציה פשוטה של בעיה אחרת מוכרת בתורת הסתברות — בעיית יום ההולדת. ההסתברות שלאדם אקראי יהיה יום הולדת באותו יום שבו לפלוני חל יום ההולדת היא 1/365. הסתברות זאת אנלוגית להסתברות ההתאמה המקרית בבדיקת ה־DNA אלא שהיא קבועה ואיננה משתנה מאדם לאדם כמו בבדיקת ה־DNA. בבעיית יום ההולדת המקורית, השאלה היא כמה אנשים צריכים להיות בקבוצה, כדי שההסתברות שימצא ביניהם לפחות זוג אנשים אחד, להם יום הולדת באותו יום, היא 0.5.[15] התשובה המפתיעה היא כי מספיק שיהיו בה 23 בני אדם.[16]

---

7 הנחה זאת רלוונטית רק לשאלת המחקר השנייה.

8 הלפרט ופרדס, לעיל ה"ש 4, בעמ' 426-425.

9 שם.

10 נתונים אמיתיים ללא שגיאות מעבדה.

11 ראה נספח טכני א' ו־ב'.

12 עבור 3 פרופילים זהים מאנשים שונים A, B, C, נקבל את הזוגות הבאים AB, AC, BC.

13 קבוצה של N פרופילים גנטיים זהים כוללת בתוכה $z = \frac{N!}{(N-2)!x2!}$ זוגות זהים.

14 בשיטת ספירה זאת של הזוגות, בה מספר הזוגות כולל גם שלישיות, רביעיות, חמשיות וכולי, המספר המקסימאלי האפשרי של זוגות יתקבל כאשר כל הפרופילים הגנטיים זהים. ערכו במקרה זה יהיה: $\bar{n} = \frac{N \text{ x } (N-1)}{2}$ (הצבת p=1 במשוואה (1)). מכאן, שמספר הזוגות האפשרי, יכול לעלות על גודל האוכלוסיה.

15 S. E. Ahmed & R. J. McIntosh, An Asymptotic Approximation for the Birthday Problem, 26(3) *Crux Mathematicorum* 151 (2000). Available at: http://journals.cms.math.ca/cgi-bin/vault/public/view/CRUXv26n3/body/PDF/page151-155.pdf?file=page151-155.

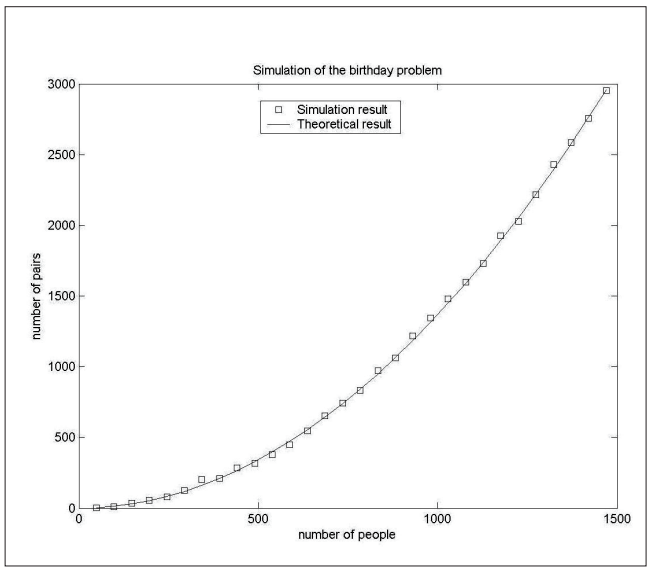
16 שם, בעמ' 154.

המחקר הנוכחי אנלוגי לא לשאלה המקורית בבעיית יום ההולדת, אלא לשאלה אחרת הקשורה להסתברות של ימי הולדת — לכמה זוגות אנשים בממוצע מקרב N אנשים, יש יום הולדת החל באותו היום? משוואה (1) נותנת את הפתרון גם לבעיה אנלוגית זאת כאשר מציבים במשוואה (1) הסתברות התאמה מקרית קבועה $\bar{p} = \frac{1}{365}$.

כדי לשכנע בתשובה פשוטה זאת, נערך במחקר סימולציה ממוחשבת בה הוגרלו N מספרים אקראיים בין 1-365, המייצגים ימי הולדת של N אנשים אקראיים ונספרו כמות התאמות בין המספרים כפונקציה של N. התוצאות הושוו לתוצאה התיאורטית המתקבלת ממשוואה 1 והן מוצגות להלן:

נציין כי ציור 1 מתאר הרצה אחת בלבד של הסימולציה. ניתן לראות כי תוצאות הסימולציה תואמות (בגבולות הסטייה הסטטיסטית) את התחזית התיאורטית. לו היינו מבצעים מספר רב של ניסויים כאלה, הממוצע של כולם היה מתכנס לתוצאה התיאורטית.

ציור 1: כמות ההתאמות בין ימי ההולדת של N אנשים כפונקציה של N



כדי להשתמש בנוסחה (1) לצורך חישוב מספר זוגות הפרופילים הגנטיים הזהים בממוצע, בקרב אוכלוסיה כגודל אוכלוסיית ישראל (N=7000000), יש צורך לדעת את הסתברות ההתאמה המקרית הממוצעת לפרופילים גנטיים. ניתן לחשב ערך זה מתוך טבלאות שכיחות אללים באוכלוסיה בהתאם לכמות האתרים שנבחרים.[17]

עבור תשעה אתרים הערך התיאורטי המתקבל הוא:

$$(2)\ \bar{p} = 6.067497 \times 10^{-10}$$

היינו, אחד חלקי 1.649 מיליארד. ומכאן , הצבת המספרים בנוסחה (1) תיתן

$$(3)\ \bar{n} = \frac{7 \times 10^6 \times (7 \times 10^6 - 1)}{2} \times 6.067497 \times 10^{-10} = 14865$$

17 ראה הסבר על דרך החישוב, בנספח טכני — ג'. הטבלאות נלקחו מ־
Motro, U., Oz, C., Adelman, R., Davidson, A., Gast, A., Hermon, D., Shpitzen, M., Zamir, A., and Freund, M. (2002) Allele frequencies of nine STR loci of Jewish and Arab populations in Israel. *Int.J.Legal Med.* 116(3): 184-186.

היינו, בקרב שבעה מיליון פרופילים גנטיים המורכבים מתשעה אתרים ואשר מתפלגים לפי שכיחויות אללים של יהודים בישראל, נצפה תחת הנחות המחקר ל־14,865 זוגות פרופילים זהים.[18]

עבור שישה אתרים[19] הערך התיאורטי המתקבל הוא:

$$(4)\ \bar{p} = 1.5379 \times 10^{-6}$$

ומכאן, הצבת המספרים בנוסחה (1) תיתן את מספר הזוגות הממוצע עבור שבעה מיליון פרופילים גנטיים בני שישה אתרים:

$$(5)\ \bar{n} = \frac{7 \times 10^{6} \times (7 \times 10^{6} - 1)}{2} \times 1.5379 \times 10^{-6} = 3.7679 \times 10^{7}$$

היינו, כ־37 מיליון ו־700 אלף זוגות תואמים.[20] תוצאה זו איננה צריכה להפתיע מאחר ואם הסתברות ההתאמה המקרית הממוצעת בסדר גודל של אחד למיליון, הרי שנצפה שיהיו התאמות רבות בין שבעה מיליון פרופילים גנטיים. כמו שנראה בסימולציה הנומרית יש גם שלישיות, רביעיות עד פרופיל גנטי אחד שמשותף ל־250 איש.

בדיקת תוצאה תיאורטית זאת תהיה אפשרית רק אם יהיו בידינו הפרופילים הגנטיים של שבעה מיליון יהודים. אנו צופים שוני בין התוצאות התיאורטיות במחקר זה לבין המציאות כתוצאה מכך שבמציאות יש קרובי משפחה, תת־אוכלוסיות, התרבות, לא הומוגנית, ושגיאות מעבדה. היינו, תוצאות אלה הן רק חסם תחתון לכמות ההתאמות שתימצא בפועל. אולם גם בהיעדר נתונים מציאותיים, ניתן לבדוק את התוצאות התיאורטיות על סמך סימולציה נומרית.

ד. הסימולציה

ד.1 כללי

הסימולציה שנעשתה בוצעה בעזרת תוכנה שנכתבה בשפת C במיוחד עבור מחקר זה. החלק הראשון בסימולציה הוא ליצור מאגר של פרופילים גנטיים, על פי טבלאות שכיחות אללים באוכלוסיה בישראל[21] על ידי מחולל מספרים אקראיים.[22] אנו מציגים תוצאות עבור האוכלוסיה היהודית בישראל.[23]

כל פרופיל גנטי במאגר מיוצג על ידי אוסף של זוגות מספרים — זוג לכל אתר ב־DNA הנבדק. אם למשל, בחרנו לבדוק מאגר פרופילים גנטיים אשר הופקו מתשעה אתרים, אזי כל פרופיל גנטי הוא תשעה זוגות של מספרים. כל זוג מספרים מייצג את שני

18 לצרכים מעשיים, לפעמים ניתן לצמצם את אוכלוסיית החשודים על ידי אינפורמציה נוספת. למשל, כאשר ידועה אינפורמציה הנוגעת לעבריין, כגון מינו. אם בחקירת עבירה נתונה, גודל האוכלוסיה בפועל אליה יכול להשתייך מבצע העבירה היא בגודל של מיליון פרטים, אזי על פי משוואה (1), כמות הזוגות התואמים הרלוונטיים יורדת באופן ריבועי, פי 49, מ־14,865 ל־303 זוגות תואמים בלבד. עניין אחר הוא כי גם כאשר הבדיקה נעשית על תשעה אתרים, לפעמים לא מתקבלות תוצאות ברורות מכל האתרים. למשל, התוצאות מתקבלות משבעה או 8 אתרים בלבד. במקרה זה, יש להביא בחישוב, את הסתברות ההתאמה המקרית המתקבלת מאותו מספר אתרים מוגבל ולא מכל תשעה האתרים.

19 ששת האתרים עבורם נעשה החישוב לאורך כל מאמר זה הם:
THO1., TPOX, CSF1PO, vWA, FESFPS, F13A01.

20 כאמור לעיל, בה״ש 14, בשיטת ספירה זאת, בה הזוגות כוללים גם שלישיות, רביעיות, חמישיות וכו׳, אין דבר חריג בכך שמספר הזוגות גדול מגודל האוכלוסיה כולה. מספר הזוגות עולה באופן ריבועי עם גודל האוכלוסיה.

21 Motro, U., Oz, C., Adelman, R., Davidson, A., Gast, A., Hermon, D., Shpitzen, M., Zamir, A., and Freund, M. (2002) Allele frequencies of nine STR loci of Jewish and Arab populations in Israel. *Int.J.Legal Med.* 116(3): 184-186.

22 מחולל המספרים האקראיים בו השתמשנו הוא גרסה בשפת התכנותc למחולל מספרים אקראיים המתואר אצל:
George Marsaglia and Arif Zaman, "Toward a Universal Random Number Generator", Florida State University Report: FSU-SCRI-87-50 (1987) available at:
www.jud10.org/AdministrativeOrders/LocalRules/RandomNumberGenerator.pdf.
גרסת שפת c, לאלגוריתמים זה, זמינה כאן:
http://local.wasp.uwa.edu.au:80/~pbourke/other/random/randomlib.c.

23 הטבלאות שם מכילות אינפורמציה על תשעה אתרים הבאים:
THO1, TPOX, CSF1PO, vWA, FESFPS, F13A01, D13, D7, D16.

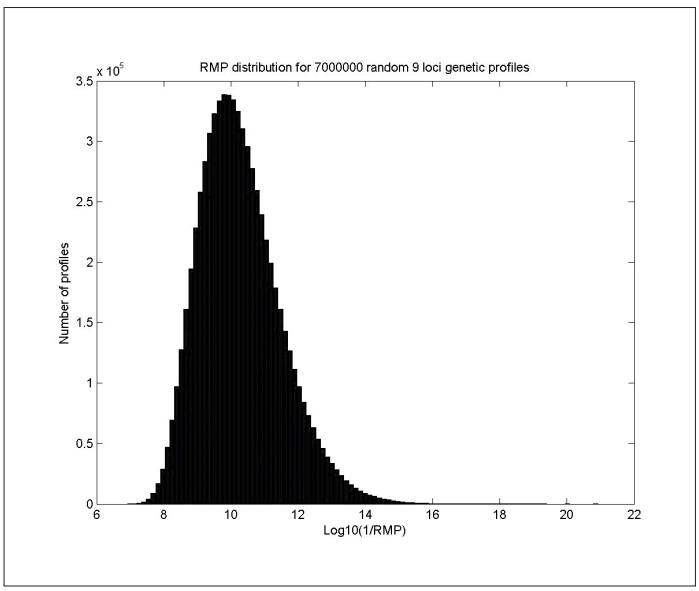
האללים שבכל אתר, כאשר המספר הראשון תמיד יהיה הקטן מביניהם. שני פרופילים גנטיים יחשבו תואמים, אם כל זוגות המספרים המופיעים בכל פרופיל, זהים. הסתברות ההתאמה המקרית חושבה על פי כלל המכפלה כאשר עבור אתרים מסוג heterozygous (בהם שני האללים שונים), הוסף גורם 2 למכפלה.[24] כדי לבדוק שלא התרחשה טעות בהליך ייצור הפרופילים הגנטיים, בדקנו אחרי יצורם, כי התפלגות האללים במאגר זהה להתפלגות האללים כפי שעולה מטבלאות שכיחות האללים באוכלוסיה.[25] הסימולציות בוצעו עבור תשעה אתרים ועבור שישה אתרים.

ספירת מספר ההתאמות בקרב שבעה מיליון פרופילים גנטיים מצריכה מספר גדול מאוד של השוואות (7000000 x 7000000). כדי לפתור את בעיית זמן הריצה ולייעל את הליך ההשוואה נעשה שימוש במחקר זה ב־Hash Table[26] באופן שיאפשר את ביצוע החישובים במחשב ביתי סטנדרטי, במשך זמן סביר. עיבוד התוצאות והגרפים בוצעו בעזרת תוכנת MATLAB.

2.ד תוצאות סימולציה – ספירת התאמות עבור שבעה מיליון פרופילים גנטיים המורכבים מתשעה אתרים המתפלגים על פי טבלאות שכיחות אללים באוכלוסיה היהודית

ראשית, להלן היסטוגרמה המציגה את כמות הפרופילים הגנטיים כפונקציה של הלוגריתמים על בסיס 10 של אחד חלקי הסתברות ההתאמה המקרית[27]:

ציור 2: היסטוגרמה של התפלגות הסתברות התאמה מקרית של 7 מיליון פרופילים גנטיים המורכבים מתשעה אתרים



24 Comm on DNA Forensic Sci: An Update, Comm'n on DNA Forensic Sci: An Update, Nat'l Research Council, The Evaluation of Forensic DNA Evidence 92 (Nat'l Acad Press, 1996).

25 בדיקה זו מראה גם כי מחולל המספרים האקראיים שהשתמשנו בו אכן בעל מחזוריות מספיק גדולה.

26 Kyle Loudon, Mastering Algorithms with C, 141-177 (1999) ISBN: 1-56592-453-3. ראו פירוט נספח טכני ד'.

27 באופן זה, הסתברות התאמה מקרית של אחד למיליארד, תוצג בציר ה־x כ־9 לפי החישוב הבא: $LOG10\left(\frac{1}{10^{-9}}\right) = 9$.
הרזולוציה בהיסטוגרמה של 100, עבור ערכי אחד חלקי הסתברות ההתאמה המקרית, לפני הוצאת ה־LOG.

ראשית, ניתן לראות מציור 2, כי טווח התפלגות הסתברות ההתאמה המקרית הוא עצום. יש פרופילים גנטיים אשר הסתברות ההתאמה המקרית שלהם אחד חלקי עשר בחזקת שבע ויש גם כאלה אשר הסתברות ההתאמה המקרית שלהם אחד חלקי עשר בחזקת 22.

שנית, הממוצע של הסתברות התאמה מקרית עבור תשעה אתרים חושב מהסימולציה והשווה לתוצאת החישוב התיאורטי הנתונה במשוואה (2). התוצאה שהתקבלה בסימולציה אכן קרובה לתוצאה התיאורטית

$$(6)\ \overline{p\text{-}simulation} = 6.067813 \times 10^{-10}$$

$$(2)\ \bar{p} = 6.067497 \times 10^{-10}$$

אחזור ואציין כי מדובר על הרצה בודדת. ממוצע ממספר רב של הרצות אמור להתכנס לתוצאה התיאורטית המדויקת, אולם עקב מגבלות זמן ההרצה לא בוצעו במחקר זה הרצות רבות כאלה.

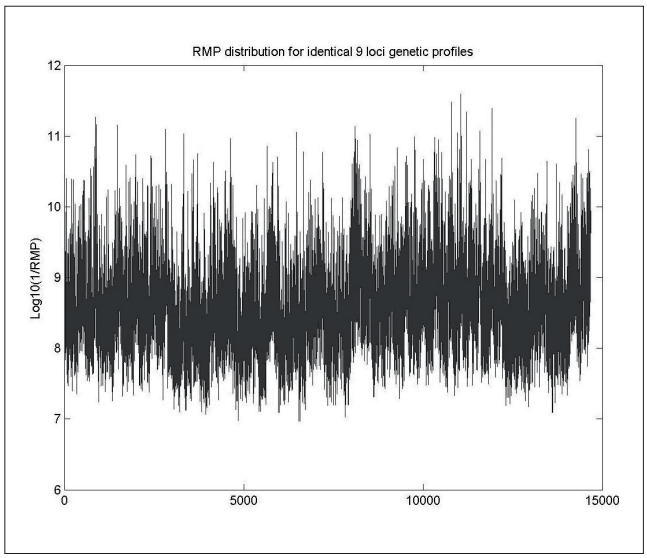
שלישית, ספירת ההתאמות הניבה תוצאות אשר דומות מאוד לתחזיות התיאורטיות:[28]

14,264 זוגות פרופילים גנטיים זהים.

212 שלישיות פרופילים גנטיים זהים (שהם 636 זוגות).

6 רביעיות פרופילים גנטיים זהים (שהם 36 זוגות).

ציור 2 (א) : לוגריתמים של אחד חלקי הסתברות ההתאמה המקרית עבור כ־14000 התאמות המתקבלות בקרב שבעה מיליון פרופילים גנטיים באוכלוסיה יהודית



28 אציין כי שיטת ספירת הזוגות בסימולציה, שונה משיטת ספירת הזוגות בתוצאה התיאורטית הנתונה במשוואה (1). ספירת הזוגות בסימולציה, כוללת רק זוגות ולא שלישיות, אשר נספרות בנפרד. באותו אופן, ספירת שלישיות כוללת רק שלישיות ולא רביעיות, אשר נספרות בנפרד. בשיטת הספירה בסימולציה, הסכום של מספר הפרופילים הגנטיים שאין להם תואם גנטי, עם מספר הזוגות המוכפל ב־2, עם מספר השלישיות המוכפל ב־3, וכו', הוא בדיוק שבעה מיליון.

הזוגות, מתקבלים גם בהסתברויות התאמה מקרית נמוכות במיוחד, עד הסתברות של אחד ל־400 מיליארד כפי שמוצג בציור לעיל:[29]

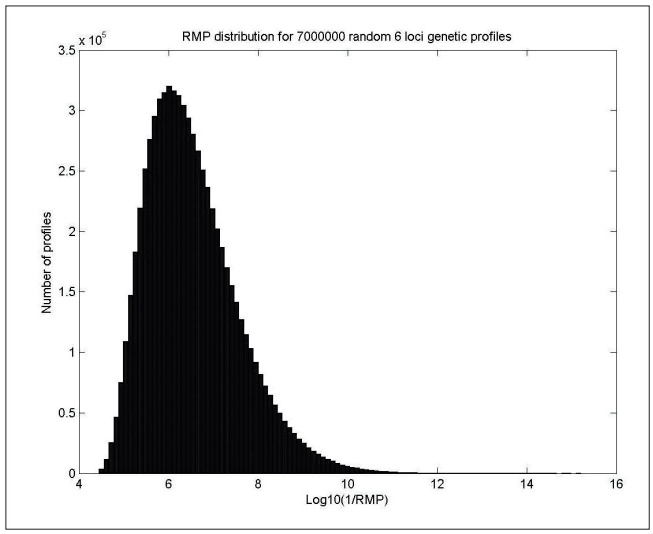
אציין כי בהרצה אחרת, התקבל פרופיל גנטי אחד המשותף לחמישה בני אדם.

ד.3 תוצאות סימולציה – ספירת התאמות עבור שבעה מיליון פרופילים גנטיים המורכבים משישה אתרים המתפלגים על פי טבלאות שכיחות אללים באוכלוסיה היהודית

ראשית, להלן היסטוגרמה המציגה את כמות הפרופילים הגנטיים כפונקציה של הלוגריתמים על בסיס 10 של אחד חלקי הסתברות ההתאמה המקרית[30]:

ניתן לראות מציור 3, כי טווח התפלגות הסתברות ההתאמה המקרית הוא עצום. יש פרופילים גנטיים אשר הסתברות ההתאמה המקרית שלהם אחד חלקי עשר בחזקת חמש ויש גם כאלה אשר הסתברות ההתאמה המקרית שלהם אחד חלקי עשר בחזקת 16.

ציור 3: היסטוגרמה של התפלגות הסתברות התאמה מקרית של שבעה מיליון פרופילים גנטיים המורכבים משישה אתרים



שנית, הממוצע של הסתברות התאמה מקרית עבור שישה אתרים חושב והשווה לתוצאת החישוב התיאורטי הנתונה במשוואה (4). התוצאה שהתקבלה בסימולציה אכן קרובה לתוצאה התיאורטית.

$$(7)\ \overline{p\text{ - }simulation} = 1.5363 \times 10^{-6}$$

$$(4)\ \bar{p} = 1.5379 \times 10^{-6}$$

29 ערכי הסתברות ההתאמה המקרית מופיעים בציר ה־y באופן לוגריתמי כמתואר בה"ש 27.

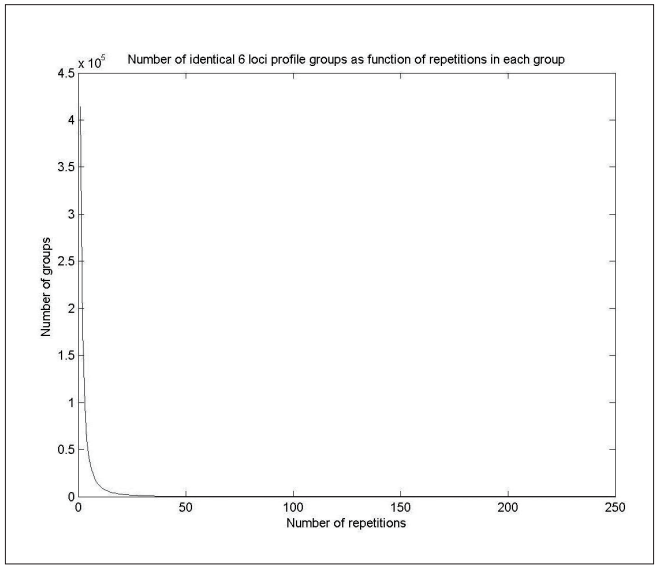
30 באופן זה, הסתברות התאמה מקרית של אחד למיליון, תוצג בציר ה־x כ־6 לפי החישוב הבא: $LOG10\left(\frac{1}{10^{-6}}\right) = 6$.

אחזור ואדגיש כי מדובר על הרצה בודדת. ערך ממוצע מהרצות רבות אמור להתכנס לתוצאה התיאורטית המדויקת, אולם עקב מגבלות של זמן הרצה לא בוצעו במחקר זה הרצות רבות כאלה.

ספירת ההתאמות הניבה כצפוי כמות גדולה של זוגות שלישיות וכו' פרופילים גנטיים תואמים ואף קבוצה של 250 אנשים החולקים פרופיל גנטי אחד כמתואר בציור הבא:[31]

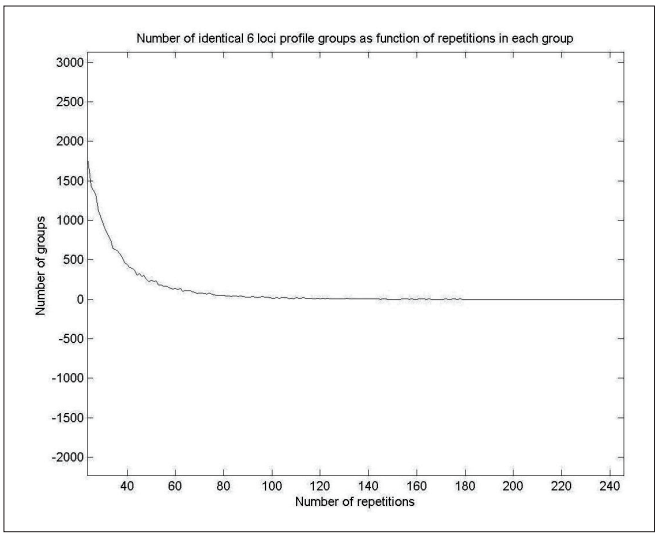
ניתן לראות מציור 4 כי יש מאות אלפים של שלישיות, רביעיות וכו' בני אדם החולקים אותו פרופיל גנטי. הגדלת הגרף (zoom), תאפשר להתמקד בקבוצות של עשרות אנשים החולקים אותו פרופיל גנטי.

ציור 4: מספר קבוצות האנשים להם פרופיל גנטי זהה, כפונקציה של מספר האנשים בקבוצה



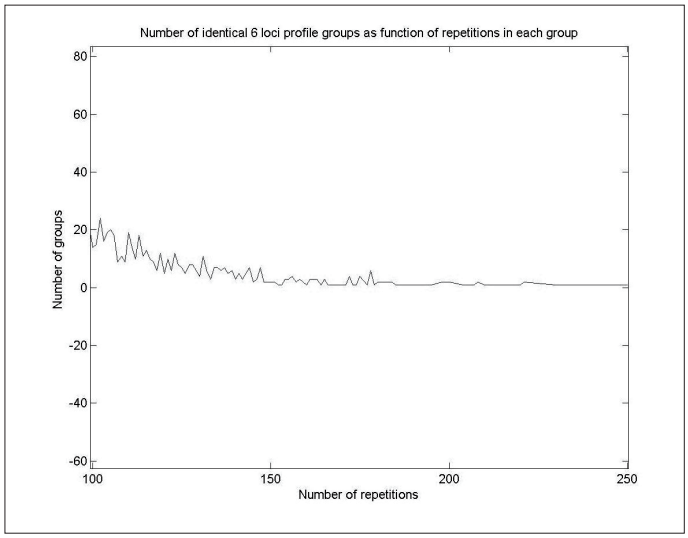
31 לגבי שיטת הספירה בסימולציה, ראו לעיל, ה"ש 28.

ציור 5: מספר קבוצות האנשים להם פרופיל גנטי זהה, כפונקציה של מספר האנשים בקבוצה.
הגדלת הגרף והתמקדות בקבוצות המכילות מעל 20 אנשים בקבוצה



ניתן לראות מציור 5 כי יש אלפי קבוצות המונות עשרות אנשים אשר חולקים אותו פרופיל גנטי. מציור 6 ניתן גם לראות כי יש קבוצה אחת אשר בה, 250 אנשים חולקים אותו פרופיל גנטי.

ציור 6: מספר קבוצות האנשים להם פרופיל גנטי זהה, כפונקציה של מספר האנשים בקבוצה.
הגדלת הגרף והתמקדות בקבוצות המכילות מעל 100 אנשים בקבוצה



ד. 4 סימולציה של חקירה משטרתית

לאחר שבדקנו בסעיף הקודם כי התוצאות התיאורטיות תואמות את אלה מהסימולציה ולאור העובדה כי קיימים באוכלוסיה פרופילים זהים, נוכל לחקור את העבודה המשטרתית למול מאגר פרופילים גנטיים. בישראל כמו בעולם כולו החלו להסתמך על מאגרי פרופילים גנטיים לצורך איתור עבריינים.[32] על פי שיטת חקירה זאת, דגימת DNA שמושארת על ידי מבצע העבירה בזירת הפשע הפלילי מושווית לדגימות במאגר. במידה ונמצאת התאמה הרי שיש בידינו חשוד. אם אכן נמצאות נגד אותו חשוד ראיות אחרות הרי שהטכנולוגיה סייעה לנו מאוד לפענח פשע. אולם יש אפשרות כי אחרי שחשוד יאותר דרך המאגר, לא תימצא ראיה כלשהי כנגדו, ודבר לא יקשר אותו לזירת הפשע הפלילי. במקרה כזה, השאלה שבית המשפט יצטרך להכריע בה, האם ראיית ה־DNA יכולה להיות ראיה יחידה להרשעה.[33]

בעולם הפורנזי העוסק בבדיקת ה־DNA, יש מחלוקת, הקשורה למידת חוזק הראיה במקרה בו ההגעה אל החשוד דרך סריקה במאגר פרופילים גנטי.[34] יש הטוענים כי הראיה נחלשת מאוד.[35] גישה זאת מכונה גישת NRC-II. ניתן להבין את הטיעון שלהם על ידי השוואה להגרלת הפיס. אם נרכוש בודד, אזי הסיכוי שהכרטיס יזכה בפרס הראשון נמוך מאוד. אולם אם נרכוש מיליון כרטיסים, הגדלנו מאוד את הסיכוי **שאחד מבין אותם מיליון כרטיסים,** יזכה בפרס הראשון. כך הדבר גם עם פרופילים גנטיים. הסיכוי להתאמה מקרית של פרופיל גנטי בודד לפרופיל הגנטי של מבצע העבירה, יכול להיות מאוד נמוך. אולם הסיכוי שהפרופיל הגנטי של עבריין שאינו נמצא במאגר, יתאים במקרה **לאחד מבין מיליון הפרופילים המצויים במאגר**, גבוה בהרבה. עמדת דוח NRC-II המבוססת על טיעון אינטואיטיבי זה, קבעה כי על המומחה בבית המשפט לדווח על הסתברות התאמה מקרית מוכפלת בגודל מאגר הנתונים.[36] הסתברות זאת מכונה בפסיקה האמריקנית, הסתברות התאמה למאגר (Database Match Probability) והיא למעשה, ההסתברות לקבל **לפחות התאמה אחת** במאגר, בהינתן שהפרופיל של מבצע העבירה אינו במאגר.[37]

---

32 פרקים ג ו־ד לחוק סדר הדין הפלילי (סמכויות אכיפה — חיפוש בגוף ונטילת אמצעי זיהוי), התשנ"ו—1996, ס"ח 1573.

33 שאלה זאת עדיין לא הוכרעה, ראו דנ"פ 9903/03 **אבו־חמאד נ' מדינת ישראל**, פדאור 04(17) 665 (2004).

34 *David J. Balding*, The DNA Database Search Controversy, 58 Biometrics 241 (2002).

35 Comm on DNA Forensic Sci: An Update, Comm'n on DNA Forensic Sci: An Update, Nat'l Research Council, The Evaluation of Forensic DNA Evidence 161 (Nat'l Acad Press, 1996) (להלן: NRC-II). ראו גם Anders Stockmarr, *Likelihood Ratios for Evaluating DNA Evidence When the Suspect is Found Through a Database Search*, 55 Biometrics 671 (1999).

36 עמדת דוח NRC-II, לעיל ה"ש 35, מתאימה רק למקרים בהם ערך המכפלה של הסתברות ההתאמה המקרית בגודל מאגר הנתונים קטן בהרבה מ־1. אחרת, מכפלה זאת תניב הסתברות גדולה מ־1, בניגוד להגדרת המושג הסתברות המקבל ערכים בין 0 ל־1 בלבד. הסבר מדויק יותר הוא ההסבר הבא: נתון כי p הוא הסתברות ההתאמה המקרית. מכאן, כי ההסתברות לא לקבל התאמה בהשוואה בודדת הוא 1-p. מכאן, כי ההסתברות לא לקבל התאמה ב־n השוואות כאשר n הוא גודל מאגר הנתונים ובהנחת אי תלות בין ההשוואות $(1-p)^n$. מכאן, כי ההסתברות לקבל **לפחות** התאמה אחת ב־n השוואות יהיה $p_n = 1 - (1-p)^n \overset{np<<1}{\approx} n \times p$. ראו הסבר מתמטי זה אצל John G. Daugman, The Importance of Being Random: Statistical Principles of Iris Recognition, *Pattern Recognition* 279, 287–288 (2003). מכאן, שההצעה בדוח NRC-II מתאימה בקירוב טוב, רק עבור מקרה בו גודל מאגר הנתונים מוכפל בהסתברות ההתאמה המקרית, קטן בהרבה מ־1. גישת הדוח נוסחה באופן כללי יותר וקוהרנטי מבחינה מתמטית, על ידי Anders Stockmarr, לעיל ה"ש 35. על פי חישוביו, יחס הנראות, **במקרה בו יישנה התאמה אחת בלבד** במאגר הנתונים, צריך להיות מחולק בגודל מאגר הנתונים. היינו, אם יחס הנראות הוא מיליארד, והמאגר מכיל מיליון פרופילים, אזי יחס הנראות אשר צריך להיות מדווח על ידי המומחה לבית המשפט הוא אלף בלבד.

37 *United States v. Jenkins*, 887 A.2d 1013,1024-1025 (Ct. App. D.C. 2005).

מהצד השני, יש הטוענים כי הראיה לא רק שלא נחלשת כתוצאה מהסריקה, אלא אף מתחזקת, אם כי באופן זניח, כל עוד גודל מאגר הנתונים אינו מתקרב לגודל האוכלוסיה כולה.[38] גישה זאת, מכונה גישת Balding & Donnelly. הטיעון שלהם מבוסס על כך שהסתברות ההתאמה המקרית, תכונה המאפיינת את נדירות הפרופיל הגנטי, ועל כן איננה קשורה לאופן בו הגענו אל החשוד. הראיה מתחזקת כאשר היא מושגת באמצעות סריקה במאגר, מאחר ובחיפוש במאגר, נשללה האפשרות כי אותם אנשים שהפרופיל הגנטי שלהם מצוי במאגר, ולא נמצאה לגביהם התאמה לפרופיל העבריין, הם מבצעי העבירה.

המחלוקת בעניין הסריקה במאגרים גנטיים נדון בפרשת Raymond Jenkins לאחר שערכאה נמוכה קיבלה את עמדת ההגנה ופסלה את הצגת ראיית ה־DNA אשר מושגת בסריקה במאגר פרופילים גנטיים, בטענה כי אין הסכמה בעולם המדעי לגבי הניתוח ההסתברותי שלה.[39] השופטים בערעור קבעו כי יש מספר שאלות הסתברותיות שונות, להן תשובות שונות.[40] השאלות הן, מהי הסתברות ההתאמה למאגר (database match probability)?[41] מהי נדירות הפרופיל הגנטי?[42] מהי המשמעות ההסתברותית לכך, שפרט לחשוד לו נמצאה התאמה, נשללה האפשרות ששאר החשודים אשר הפרופיל שלהם מצוי במאגר, הם מבצעי העבירה?[43] מאחר ובית המשפט האמריקני השתכנע כי אף לא אחד מהצדדים, לא טעה בחישוביו, והגיע למסקנה כי המחלוקות אינה על מדע אלא על רלוונטיות.[44] בית המשפט קבע, כי מה יחשב ומה לא ייחשב לרלוונטי, אינו עניין למדענים אלא לבית המשפט להכריע בו.[45] לכן, בית המשפט התיר את הצגת הראיה בהליך המשפטי. מאידך, בית המשפט האמריקני, לא נתן

38 David J. Balding & Peter Donnelly, Evaluating DNA Profile Evidence When the Suspect is Identified Through a Database Search, 41 J. *Forensic Sci*. 603 (1996). ראו גם: David J. Balding, Errors and Misunderstandings in the Second NRC Report, 37 *Jurimetrics* J. 469, 470–473 (1997); Peter Donnelly & Richard D. Freidman, DNA Database Searches and the Legal Consumption of Scientific Evidence, 97 *Mich. L. Rev*. 931 (1999); A.P. Dawid, Comment on Stockmarr's "Likelihood Ratios for Evaluating DNA Evidence, When the Suspect is Found Through a Database Search", 57 *Biometrics* 976 (2001).

על פי גישה זאת, בהינתן הנחת הומוגניות האשמה (היינו, כל אחד מהאוכלוסיה יכול להיות העבריין במידה שווה), ובהינתן כי הראיה מושגת תוך סריקת מאגר בגודל n, ובהינתן כי גודל האוכלוסיה כולה הוא N, אזי משקל הראיה גדל יחסית למקרה בו הראיה הושגה ללא סריקה, בקבוע מכפלה K הנתון במשוואה הבאה: $K = \frac{N-1}{N-n}$ ניתן לראות כי כאשר $n=N/2$, היינו כאשר המאגר מכיל חצי מהאוכלוסיה כולה, הראיה מתחזקת רק בקבוע שערכו קרוב ל־2. מכאן ההצדקה בגישה זאת, כי לצרכים מעשיים, בו המאגר קטן משיעור חצי האוכלוסיה, ניתן להתייחס למקרה הסריקה כמו למקרה הרגיל. ראו פירוט, אצל Donnelly & Freidman, בעמ' 979-984 (Appendix).

39 ראו Jenkins, לעיל ה"ש 37.

40 שם, בעמ' 1025-1024:

The rarity statistic, the database match probability, and the Balding-Donnelly formulation do not purport to address the same issue. In reality, each formula answers a distinctly different question that may be of concern in a cold hit case. As the government correctly states, the rarity statistic simply answers the question: "How rare is this specific combination of genetic material"? The database match probability answers the question: "What is the chance/probability of obtaining a match by searching this particular database"? And the Balding-Donnelly calculation answers the question: "What is the **[*1025]** chance/probability that the person identified is the source of the sample in light of the fact that all other persons in the database search were eliminated"? None of the questions are the same; more importantly, none of the answers are mutually exclusive.

41 שם.

42 שם.

43 שם.

44 שם, בעמ' 1024:

Dr. Krane, Mr. Jenkins' own expert, admitted that two issues may arise when obtaining a match from a database: (1) the rarity of the DNA profile, and (2) the probability of obtaining a match. Dr. Krane further admitted that not only could a jury conceivably want to know both numbers, but that there was nothing controversial about the science used to calculate the rarity statistic and that an initial database search does not change the rarity of a particular profile. Dr. Krane testified that he would be more than capable of calculating both rarity and database match probability, and would be able to explain and distinguish the two numbers for a jury. Dr. Krane, however, is of the belief that in a cold-hit case, the database match probability was "the question to be addressed" to the exclusion of others. In other words, Dr. Krane believes that the database match probability is more relevant than the rarity statistic.

45 שם, בעמ' 1025:

הנחיות כיצד יש להעריך את משקל ראיית ה-DNA כאשר היא מושגת בסריקה במאגר נתונים, ואיזה סטטיסטיקה היא הרלוונטית.

מאמר זה אינו דן בסוגיה זאת ואינו מנסה ליישב בין העמדות השונות. כמו כן, אין מאמר זה מנתח את פסק הדין בעניינו של Raymond Jenkins. אין חולק כי החברה תצא נשכרת במקרים בהם הפרופיל הגנטי של מבצע העבירה מצוי במאגר, וכאשר לא התרחשה שגיאת מעבדה בהפקת הפרופיל הגנטי. תחת תנאים אלה, סריקה במאגר תעזור מאוד בפענוח פשעים. ככל שהמאגר יהיה גדול יותר, כך ההסתברות שהפרופיל של מבצע העבירה בתוך המאגר גדולה יותר וכך נפענח יותר פשעים. אולם גם אם פשעים רבים יפוענחו, עדיין עומדת לנאשם חזקת החפות. ייתכן כי הפרופיל הגנטי של מבצע העבירה, אינו נמצא במאגר, ומקור ההתאמה לנאשם הוא במקריות. אפשרות זאת רלוונטית במיוחד כאשר ראיית ה-DNA ראיה יחידה וכאשר יש לפעמים ראיות אחרות לטובת החשוד.[46] לכן, יש לבחון אם האפשרות שאדם שהפרופיל הגנטי שלו מופיע במאגר גדול, יופלל כתוצאה מפשע שביצע אדם שהפרופיל הגנטי שלו אינו נמצא במאגר, אפשרות ריאלית לאורך עבודה חקירה משטרתית שוטפת של פשעים רבים אל מול מאגר פרופילים גדול.

לצורך בדיקת אפשרות זאת, ביצעתי סימולציה של עבודה משטרתית מול מאגר פרופילים גנטיים המכילים מיליון פרופילים גנטיים המורכבים משישה ותשעה אתרים שהוגרלו על פי טבלאות שכיחות אללים באוכלוסיה היהודית. הסימולציה בוצעה באופן הבא: מגרילים פרופיל גנטי אשר מדמה פרופיל גנטי של עבריין שאינו נמצא במאגר, ואשר נמצא בזירת פשע פלילי מספר 1.[47] סורקים את מאגר הנתונים ובודקים אם יש התאמה. אם נמצאה התאמה, מפסיקים וקובעים כי בניסוי מספר 1, התקבלה התאמה מקרית[48] כבר אחרי סריקה אחת. אם לא התקבלה התאמה מגרילים פרופיל גנטי שני, אשר מדמה פרופיל של עבריין אחר בפשע 2. אם נמצאה התאמה, מפסיקים וקובעים כי בניסוי מספר 1, התקבלה התאמה מקרית אחרי 2 סריקות. אם לא נמצאה התאמה ממשיכים עוד ועוד, עד שמתקבלת התאמה (למשל אחרי Y=18 ניסיונות). במקרה כזה קובעים כי בניסוי מספר 1, התקבלה התאמה מקרית ראשונה אחרי Y סריקות (בדוגמא 18). זהו ציר ה-Y בגרף — "הסריקה בה התקבלה התאמה ראשונה". מאחר ואנו עוסקים במספרים אקראיים והתוצאות בכל ניסוי יהיו אחרות, אנו עושים 500 ניסיונות כאלה. התוצאות מוצגות באופן שציר ה-X לא יהיה מספר הניסוי האקראי אלא לוג אחד חלקי הסתברות ההתאמה המקרית באותו ניסוי בו התקבלה התאמה מקרית. כך נוכל לקבל מידע לא רק לגבי ההתאמה אלא גם לגבי הסתברות ההתאמה המקרית באותה התאמה מקרית.

What is and is not relevant is not appropriately decided by scientists and statisticians. This court recognizes that as jurists we are not always in a position to determine what is good science and what is bad science. Frye directs us to defer to the determinations of the experts in the field to answer that question. Questions of relevancy, however, have never been outside of judicial competence. Determining what evidence is and is not relevant is a hallmark responsibility of the trial judge and that responsibility is not appropriately delegated to parties outside the court.

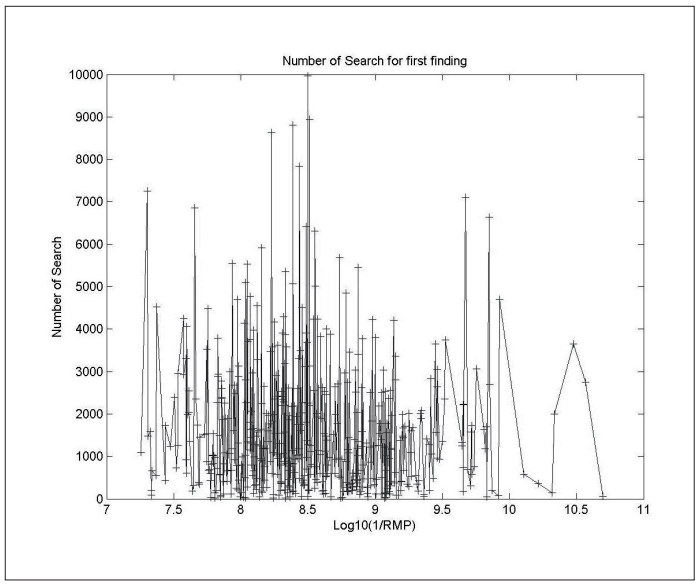
46 מקרה כזה הוא מקרה Adams באנגליה. *R. v. Denis Adams* [1996] 2 Cr. App. R. 467; *R. v. Denis Adams* (No. 2), [1998] 1 Cr. App. R. 377; Michael Lynch & Ruth McNally, "Science," "Common Sense," and DNA Evidence: A Legal Controversy about the Public Understanding of Science, 12 *Pub. Understanding Sci.* 83 (2003). במקרה זה, Denis John Adams, אשר בעת מעצרו היה בן 37, הואשם באונס על סמך ראיית DNA יחידה כאשר שכיחות הפרופיל הגנטי באוכלוסיה הוערך על ידי מומחה התביעה כאחד למאתיים מיליון (שם בעמ' 84, 89). הראיה הושגה, לאחר שנלקחה מ-Adams דגימת דם לצורך חקירת עבירה מין אחרת (שם בעמ' 84). סריקה במאגר פרופילים גנטיים הניבה התאמה לפרופיל של עבריין מקרה אונס לא מפוענח שהתרחש שנתיים לפני כן (שם, בעמ' 84). הראיה היחידה כנגד Adams הייתה אותה ראיית DNA. המתלוננת טענה כי האנס היה בחור צעיר בגילאי 20-25. היא לא זיהתה את הנאשם במסדר זיהוי. היא טענה כי הוא אינו דומה לאנס. לדבריה, Adams נראה מבוגר בהרבה מהאנס אשר היה צעיר (שם בעמ' 87). בנוסף, Adams סיפק אליבי לזמן האונס אשר נתמך בעדות חברתו. האליבי לא הופרך במשפט. למרות ראיות אלה לזכותו, Adams הורשע במשפט ובמשפט חוזר.

47 ברמה העקרונית, הדרך הנכונה לבצע סימולציה זאת, היא על ידי בניית קבוצת פרופילים נוספת, מלבד המאגר המשטרתי, אשר בה יאוכסנו כל הפרופילים באוכלוסיה שאינם במאגר המשטרתי. אחרי בניית הקבוצה הנוספת, יש לבחור ממנה באופן אקראי, פרופיל המדמה פרופיל גנטי של עבריין שאינו נמצא במאגר המשטרתי. אם לא נעשה כך, אנו מניחים באופן סמוי ושגוי, שיש מספר אין סופי של עבריינים שאינם במאגר. ברמה המעשית, מאחר וכמות הפרופילים הגנטיים המקסימאלית שהוגרלו בניסוי בודד לא עלתה על 10000, ומאחר וסך כל הפרופילים הגנטיים שהוגרלו בכל הניסויים ביחד היה פחות משמונה מאות אלף, מספר הקטן בהרבה מגודל האוכלוסיה כולה במדינת ישראל, ניתן היה לבצע את הניסוי, כפי שמתואר בגוף המאמר.

48 התאמה מקרית משמעותה כי אין מדובר בעבריין אלא באדם אחר שבאופן מקרי, הפרופיל הגנטי שלו זהה לעבריין.

ד.4. 1) תוצאות סימולציה של חקירה משטרתית עבור מאגר פרופילים גנטיים המורכבים מתשעה אתרים

ציור 7: מספר הסריקות הנדרש לקבלת התאמה מקרית כפונקציה של הסתברות ההתאמה המקרית במאגר המונה מיליון פרופילים גנטיים המורכבים מתשעה אתרים

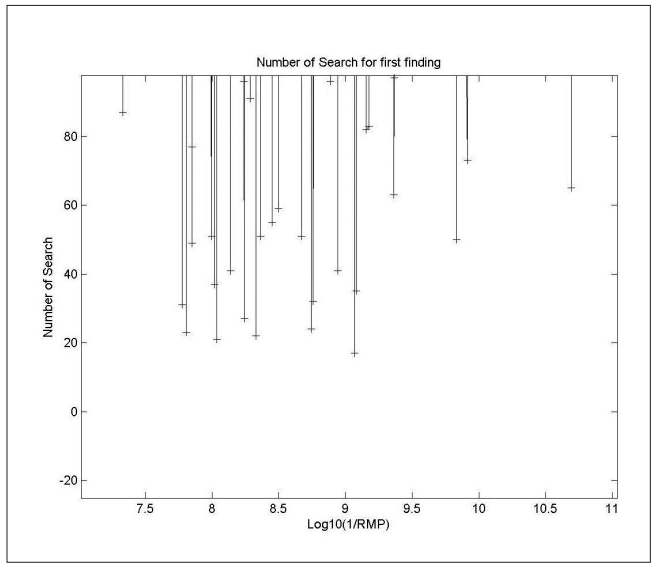


הנקודות בציור 7 חוברו, כדי שהקורא יוכל לחוש את האופי הסטוכסטי של ההתאמות המקריות. היינו, התאמה מקרית יכולה להתקבל אחרי עשרת אלפים סריקות וגם אחרי פחות מעשרים סריקות. אנו רואים בציור 7, כי התאמה מקרית יכולה להתקבל גם כאשר הסתברות ההתאמה המקרית נמוכה עד כדי אחד ל־50 מיליארד (אחד חלקי 10 בחזקת 10.7).

כדי לראות את המקרים בהם התאמה מקרית התקבלה לאחר מספר קטן של סריקות הגרף הוגדל כך שהוא מתמקד רק באותן התאמות מקריות שהתקבלו לאחר פחות מ־100 סריקות.

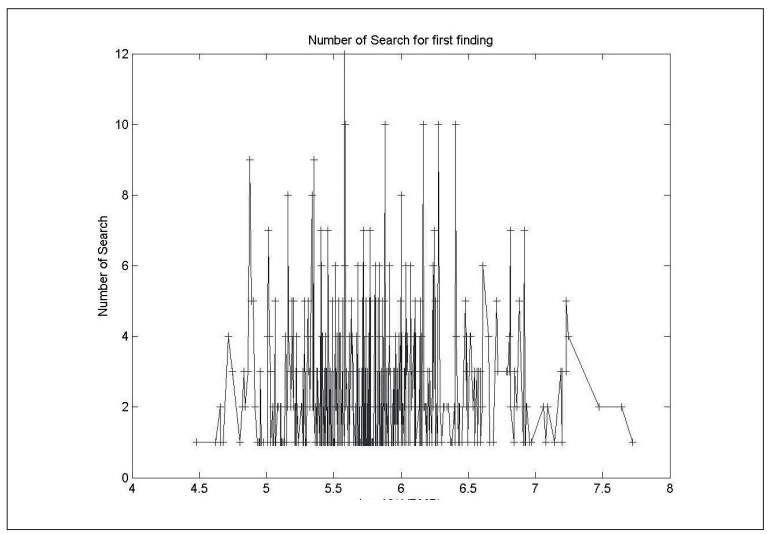
ניתן לראות מציור 8, כי התאמה מקרית יכולה להתקבל אחרי כ־65 סריקות, כאשר הסתברות ההתאמה המקרית נמוכה עד כדי אחד ל־50 מיליארד (אחד חלקי 10 בחזקת 10.7). התאמה מקרית יכולה להתקבל עבור פרופיל גנטי שהסתברות ההתאמה המקרית שלו אחד למיליארד, אחרי פחות מ־20 סריקות. כל התאמה מקרית כזאת משמעותה הרשעת חף מפשע, אם יסתמכו על אותה התאמה ולא יידרשו לראיות סיוע. לסיכום האמור לגבי פרופילים גנטיים המורכבים מ־9 אתרים, ניתן לראות כי מסוכן להסתמך רק עליהם לצורך הרשעה. התאמה מקרית אפשרית גם לאחר פחות 20 סריקות, היינו, לאחר 20 פשעים אותם המשטרה חוקרת ובהם הפרופיל של העבריין האמיתי אינו נמצא במאגר.

ציור 8: מספר הסריקות הנדרש לקבלת התאמה מקרית כפונקציה של הסתברות ההתאמה המקרית במאגר המונה מיליון פרופילים גנטיים המורכבים מתשעה אתרים –
הגדלה (zoom) והתמקדות במספר סריקות נמוך מ-100



ד.4. 2) תוצאות סימולציה של חקירה משטרתית עבור מאגר פרופילים גנטיים המורכבים משישה אתרים

ציור 9: מספר הסריקות הנדרש לקבלת התאמה שגויה כפונקציה של הסתברות ההתאמה המקרית במאגר המונה מיליון פרופילים גנטיים המורכבים משישה אתרים



ניתן לראות שהתאמות שגויות עלולות להתקבל כבר בסריקה ראשונה וגם בהסתברויות התאמה מקרית של אחד לשישים מיליון. מכאן, כי בכל מקרה בו פשע בוצע על ידי עבריין מחוץ למאגר, יופלל בסבירות גבוהה אדם שהפרופיל הגנטי שלו מצוי בתוך המאגר.[49]

ה. דיון בתוצאות והמלצות

בספרות העוסקת בתחום נטען כי אין להרשיע אדם על סמך ראיית DNA לבדה, וזאת בעיקר לאור האפשרות של שגיאת מעבדה.[50] תוצאות מחקר זה מלמדות על סיבה נוספת והיא כי גם הסתברויות התאמה מקרית של אחד למיליארדים אינן מספקות לקביעת ייחודיות הפרופיל הגנטי. המחקר מראה בוודאות כי גם שכיחות של אחד למיליונים אינה מספקת. בנסיבות בהן תיתכן שגיאה בבדיקת ה-DNA או כתוצאה משגיאת מעבדה או התאמה מקרית ובנסיבות בהן התאמה בבדיקה תספיק להרשעה, כל אדם שהפרופיל הגנטי שלו נמצא במאגר, הוא חשוד תמידי שאינו יכול שלא לחשוש כי בעוד שנה או שנתיים ואף עשר שנים הוא יואשם על סמך התאמה בבדיקת DNA, בעבירה אותה הוא לא ביצע, ואשר ייתכן שבוצעה גם שנים רבות קודם לכן.[51]

ב-1999 בוצעה באנגליה סריקה במאגר פרופילים גנטיים כדי לאתר חשוד בפריצה.[52] הסריקה העלתה כי דגימת ה-DNA שנמצאה בזירת הפשע שייכת לאדם מסוים בהסתברות התאמה מקרית של אחד לשלושים ושבעה מיליון. אותו אדם, גר 200 מייל מזירת הפשע, חולה במחלת פרקינסון ואף אינו מסוגל לנהוג. הוא גם סיפק אליבי לזמן הפשע. הסתברות ההתאמה המקרית התקבלה כתוצאה מבדיקת 6 סמנים בפרופיל הגנטי. רק בגלל התעקשות סניגורו של הנאשם, כאשר בדקו 4 סמנים נוספים, התברר כי אין התאמה בבדיקת ה-DNA. לאור התוצאות שהראינו במחקר זה, אין הדבר מפתיע כלל ועיקר.

מכאן שיש לעשות כל שניתן, כדי להקטין את הסכנה בהרשעה על סמך בדיקת DNA. מאמר זה מתמקד באפשרות השגיאה הגנטית ובדרך לצמצומה. שגיאה זאת אינה גזירת גורל. אם נרצה שמספר הזוגות הממוצע בכל כדור הארץ יהיה נמוך מערך סף מסוים (נניח $\bar{n} = 0.01$) נוכל לחשב את הסתברות ההתאמה המקרית הממוצעת הנדרשת לכך. לצורך חישוב זה, ניתן לרשום את משוואה (1) על ידי בידוד הסתברות ההתאמה המקרית הממוצעת והצבת המספר שבעה מיליארד, עבור N, כמספר בני האדם החיים על פני כדור הארץ:

$$(8)\ \bar{p} = \frac{2 \times \bar{n}}{N \times (N-1)} = \frac{2 \times 0.01}{7 \times 10^9 \times (7 \times 10^9 - 1)} = 4.08 \times 10^{-22}$$

---

49 סביר מאוד כי יהיו מקרים רבים בהם תימצא יותר מהתאמה אחת בתוך המאגר. אולם גם אם נשללו שאר החשודים להם נמצאה התאמה במאגר בראיות חיצוניות, ונשארנו עם חשוד אחד, הדבר אמור להדליק תמרור אזהרה שכן כמו שבמאגר נמצאו מספר התאמות, כך גם מחוץ למאגר עלולים להיות עוד מספר רב של אנשים אשר גם הפרופיל הגנטי שלהם תואם את זה של מבצע העבירה.

50 ראו Sangero and Halpert לעיל, ה"ש 4. ראו גם הלפרט ופרדס, שם. ראו גם, יורם פלוצקי "משקלה של ראיית DNA – בעקבות פסק הדין מוראד אבו חמאד" **רפואה ומשפט** 30, 174, 178 (2004).

51 אפילו קורבנות עבירה פלילית אשר הפרופיל הגנטי שלהם היה במאגר לצרכי חקירת פשע שבוצע נגדם אינם יכולים שלא לחשוש. כך קרה בחקירת רצח הפעוטה Jaidyn Leskie. התאמת DNA מזירת רצח הפעוטה הובילה לנערה מפגרת שנאנסה, אשר הפרופיל הגנטי שלה הופק לצרכי חיפוש זה שאנס אותה. חקירת המקרה והעובדות המסוימות בו, הובילו לניקוי החשדות כנגד אותה נערה ומקור השגיאה הוסבר על ידי מומחים שבדקו את המקרה, בזיהום הדגימות. הסבר זה נראה סביר, מכיוון ששתי הדגימות עובדו באותה מעבדה בערך באותה תקופת זמן. William C. Thompson, Tarnish on the "Gold Standard": Understanding Recent Problems in Forensic DNA Testing, 30 *Champion* 10, 13–14 (2006). במקרה אחר, ב-Michigan שבארה"ב נחקר מחדש רציחתה של הסטודנטית Jane Mixer משנת 1969, בעזרת טכנולוגיית DNA אשר לא הייתה זמינה בחקירת הרצח המקורית. החקירה המחודשת הובילה להתאמה לשני חשודים, ל-John Ruelas ול-Gary Leiterman. אלא ש-John Ruelas היה בן ארבע בזמן הרצח ולכן הוברר כי לא הוא הרוצח. מנגד, כן הוגש כתב אישום כנגד Gary Leiterman. למרות צירוף המקרים של עובדת חוסר הסבר הגיוני להתאמה בין דגימות הרוצח לדגימה של John Ruelas עם העובדה כי גם במקרה זה הוברר כי שלושת הדגימות, זאת של הקורבן ושל שני החשודים, עובדו באותה מעבדה באותו זמן, Leiterman, הורשע ברצח הסטודנטית. שם, בעמ' 14.

52 Jennifer L. Mnookin, Fingerprint Evidence in the Age of DNA Profiling, 67 *Brook. L. Rev.* 13, 50-51 (2001).

אם ונסתפק בייחודיות הפרופיל הגנטי בקרב תושבי מדינת ישראל בלבד (המונים כשבעה מיליון תושבים) ובאותו סף, תידרש הסתברות ההתאמה המקרית הממוצעת:

$$(9)\ \bar{p} = \frac{2 \text{ x } \bar{n}}{N \text{ x } (N - 1)} = \frac{2 \text{ x } 0.01}{7 \text{ x } 10^6 \text{ x } (7 \text{ x } 10^6 - 1)} = 4.08 \text{ x } 10^{-16}$$

הסתברויות התאמה מקרית נמוכות אלה יתקבלו בטכנולוגיה של ימינו, אך ורק אם נגדיל את כמות האתרים בבדיקת ה-DNA. מחקר זה אינו יכול להמליץ על כמות האתרים הנדרשת להשגת יעד זה מאחר ויש חשיבות לבחירת אתרים בהם השונות בין בני האדם גדולה. כדי להשתמש במספר רב יותר של אתרים, יש צורך גם במחקר שיקבע את שכיחות האללים באוכלוסיות הנמצאות בישראל, באותם אתרים נוספים שישמשו לביצוע הבדיקה.

מחקר זה גם מאפשר לנבא מתי צפויה התאמה כפולה ראשונה במאגר פרופילים גנטיים. לצורך כך, נרשום את משוואה (1) קצת אחרת:

$$(10)\ \bar{n} = \frac{N \text{ x } (N - 1)}{2} \text{ x } \bar{p} \Rightarrow \frac{2 \text{ x } \bar{n}}{\bar{p}} = N \text{ x } (N - 1) \overset{1<<N}{\approx} N^2$$

מכאן נקבל ש:

$$(11)\ N \approx \sqrt{\frac{2 \text{ x } \bar{n}}{\bar{p}}}$$

אם נציב במשוואה (11) את הערך עבור מספר הזוגות הממוצע $\bar{n} = 1$ ואת הערך התיאורטי ממשוואה (2) עבור הסתברות של התאמה מקרית ממוצעת של פרופילים בני 9 אתרים $\bar{p} = 6.067497 \text{ x } 10^{-10}$, נוכל להעריך את גודל מאגר הנתונים הממוצע בו אנו צפויים למצוא כפילות ראשונה במאגר הנתונים. ערך זה הוא:

$$(12)\ N \approx \sqrt{\frac{2 \text{ x } 1}{6.067497 \text{ x } 10^{-10}}} = 57413$$

מכאן, שאנו מצפים לקבל בממוצע זוג ראשון תואם במאגר, לאחר שהוא יכיל כ-57,413 פרופילים גנטיים בעלי 9 אתרים.

חישוב דומה עבור 6 אתרים ושימוש בערך ממשוואה (4) עבור הסתברות התאמה מקרית ממוצעת בפרופילים גנטיים בעלי 6 אתרים, $\bar{p} = 1.5379 \text{ x } 10^{-6}$ ייתן כי

$$(13)\ N \approx \sqrt{\frac{2 \text{ x } 1}{1.5379 \text{ x } 10^{-6}}} = 1140$$

היינו, אנו מצפים כי כבר בקרב 1,140 פרופילים המורכבים מ-6 אתרים המצויים במאגר, נגלה זוג אחד תואם משני אנשים שונים.

מכאן החשיבות המחקרית בנתוני המאגר המוחזקים על ידי הרשויות. אלה אינן תיאוריות או סימולציות, כי אם נתונים מציאותיים שניתן להשוותם לתיאוריה, ובכך הם יכולים להוות מבחן הפרכה נוסף לתיאוריה הגנטית והפורנזית.[53] מחקר זה ממליץ לרשויות למצוא את הדרך, אשר מחד תשמור על פרטיות האנשים אשר הפרופיל הגנטי שלהם מצוי במאגר ומנגד, תאפשר חקירת המאגר ובדיקת התיאוריה.

נספח טכני

א. חישוב מספר הזוגות להם יום הולדת באותו יום, מקרב N אנשים

נסמן את כמות ההשוואות אותה יש לבצע כדי לקבל תשובה לשאלה. יום הולדתו של כל אדם מקרב N האנשים אמור להיות מושווה ליום הולדתם של $N\text{-}1$ אנשים אחרים. מכאן, שמספר ההשוואות הוא $\frac{N \text{ x } (N-1)}{2}$. החילוק ב־2 הכרחי כדי למנוע ספירה כפולה.[54] את מספר ההשוואות יש להכפיל בהסתברות קבלת התאמה מקרית בכל השוואה. מכאן שאם $p=1/365$ הוא הסתברות התאמה מקרית אזי $\frac{N \text{ x } (N-1)}{2} \text{ x } p$ $= \bar{n}$ הוא מספר הזוגות הממוצע להם יש יום הולדת באותו יום.

ב. חישוב מספר הפרופילים הגנטיים התואמים מבין N פרופילים

כדי לדעת כמה פרופילים גנטיים באוכלוסיה שגודלה N, תואמים בממוצע לפרופיל גנטי I, המאופיין בהסתברות התאמה מקרית, p_i, יש צורך לערוך $N\text{-}1$ השוואות ולהכפיל כל אחת בהסתברות ההתאמה המקרית — $(N-1) \text{ x } p_i$. מאחר והאוכלוסיה מונה N פרופילים יש לסכם על כל האוכלוסיה ולחלק ב־2 כדי למנוע ספירה כפולה. ביצוע פעולות אלה מוכיח את משוואה 1.

$$\bar{n} = \frac{1}{2}\sum_{i=1}^{N}(N-1) \text{ x } p_i = \frac{1}{2}(N-1)\sum_{i=1}^{N} p_i = \frac{1}{2}(N-1) \text{ x } N\sum_{i=1}^{N}\frac{p_i}{N} = \frac{N \text{ x } (N-1)}{2} \text{ x } \bar{p}$$

ג. חישוב הסתברות התאמה מקרית ממוצעת מטבלאות שכיחות אללים באוכלוסיה

ראשית נחשב הסתברות התאמה מקרית ממוצעת באתר אחד מסוים. באתר אחד יש 2 אללים היכולים לקבל n ערכים שונים בהתאם לטבלאות שכיחות אללים באוכלוסיה. נסמן ב־p_i את ההסתברות להימצאות אלל מסוים באתר מסוים. נסמן ב־n את מספר האללים האפשריים באותו אתר כאשר $1 \le i \le n$. נסמן $p_{i,j} = \begin{cases} 2 \text{ x } p_i \text{ x } p_j \text{ if i} \neq \text{j} \\ p_i^2 \text{ if i} = \text{j} \end{cases}$ ונסמן ב־$\bar{p}(k)$ את הסתברות ההתאמה המקרית הממוצעת של אתר מסוים — k. מכאן נקבל כי

$$\bar{p}(k) = \frac{\sum_{i,j,i\le j}^{n} p_{i,j}^2}{\sum_{i,j,i\le j}^{n} p_{i,j}}$$

---

53 על פי הפילוסוף קארל פופר, תיאוריות מדעיות לא ניתן להוכיחן. ניתן רק לנסות להפריכן. ככל שהתיאוריה עומדת בהצלחה ביותר מבחני הפרכה כך היא נחשבת מדעית יותר..Karl R. Popper, *Conjectures And Refutations: The Growth Of Scientific Knowledge* 33-35 (1969) עיקרון ההפרכה של קארל פופר התקבל בהלכת *Daubert* כאחת הדרישות למדעיות הראיה: *Daubert v. Merrell Dow Pharm*., Inc., 509 U.S. 579 (1993); ראו גם יונתן דייויס "קבילות ומשקל ראיות מדעיות במשפט — האם יש לייבא את הלכת דאוברט?" **רפואה ומשפט** 29, 50 (2003).

54 מספיק להשוות את יום הולדתו של A ליום הולדתו של B ואין להשוות גם את יום הולדתו של B עם יום הולדתו של A.

הגודל $\bar{p}$ שחושב הוא הסתברות התאמה מקרית של אתר אחד מסוים. מאחר וממוצע של מכפלה שווה למכפלת הממוצעים — $\overline{x \times y} = \bar{x} \times \bar{y}$, כל שנותר הוא להכפיל את הסתברויות ההתאמה המקרית הממוצעות מכל אתר ואתר. החישוב עצמו בוצע בתוכנה.

ד. שימוש ב-Hash Table

דרך לא יעילה להשוות בין N פרופילים ל-N פרופילים אחרים היא על ידי ביצוע של $\frac{N \times (N-1)}{2}$ פעולות השוואה ישירות. כדי להוריד את כמות החישובים נחשוב על מצב בו ניתן לחלק את הפרופילים ל-k קבוצות זרות[55], השוות בגודלן והמונות $\frac{N}{k}$ פרופילים כל אחת. במקרה זה כמות החישובים תהיה:

$$k \times \frac{\frac{N}{k} \times \left(\frac{N}{k} - 1\right)}{2} = \frac{1}{k} \times \frac{N \times (N - k)}{2}$$

בצורה זאת, כמות החישובים פחתה לפחות פי $\frac{1}{k}$. למעשה, כל חילוק לקבוצות זרות, גם אם הן אינן זהות בגודלן יפחית את זמן הריצה. במחקר זה נעשה שימוש בעקרון זה כאשר הפרופילים הגנטיים נשמרו בקבוצות זרות, בהתאם לאללים שלהם במספר חלקי של אתרים מכלל האתרים בפרופיל.[56] המימוש בתוכנה בוצע בעזרת Hash Table. המפתחות (keys) ל-Hash Table היו הפרופילים הגנטיים. פונקציית ה-Hash שמומשה במחקר זה, העניקה לכל פרופיל גנטי אינדקס (מספר שורה בטבלה) על פי האללים באותו מספר חלקי של אתרים שנבחרו. בצורה זאת, האינדקס המקסימאלי שפרופיל גנטי עשוי לקבל הוא המספר המקסימאלי של פרופילים גנטיים שניתן ליצור מאותו מספר חלקי של אתרים הוא גם גודל ה-Hash Table.[57] לצרכי המחקר המסוים הזה, צורת מימוש זאת, הספיקה כדי לעמוד במגבלות של זיכרון וזמן ריצה סביר במחשב ביתי סטנדרטי.

הערות למאמר:

1. זיהוי אישי באמצעות פרופיל גנטי מתבסס על הנחת העבודה הבסיסית כי אין שני אנשים (למעט תאומים זהים) אשר יש להם מולקולת DNA זהה. לפיכך אם השוואת דגימת ה-DNA מהזירה עם ה-DNA של החשוד הייתה מניבה זהות מלאה בכל רצף מולקולת ה-DNA, לא היה כל בסיס למחלוקת או להתניות הסתברותיות לקביעה כי מקור ה-DNA בזירה הוא בחשוד.

2. היות והשוואת הרצף המלא של מולקולת ה-DNA אינה אפשרית מבחינה מעשית, מבצעים השוואה בין הפרופיל הגנטי של הדגימה מהזירה לבין הפרופיל הגנטי של החשוד. הפרופיל הגנטי הוא מדגם של מולקולת ה-DNA המלאה, מהתוצאות המתקבלות מהבדיקה הנעשית על המדגם מסיקים מסקנות על מולקולת ה-DNA השלמה.
הגישה המחקרית או היישומית, באמצעותה חוקרים שאלה באמצעות מדגם ומיישמים את המסקנה על כל האוכלוסיה כולה, אינה ייחודית לזיהוי אישי באמצעות פרופיל גנטי ונעשית כדבר שבשגרה בתחומים מחקריים וישומיים רבים.

---

55 הכוונה במלה זרות כי פרופיל מסוים מקבוצה א לא יכול להיות זהה לאף אחד מהפרופילים בכל הקבוצות האחרות. באופן זה מספיק להשוות את הפרופילים הגנטיים בכל קבוצה בנפרד ולא לבצע השוואות בין פרופילים הנמצאים בקבוצות שונות.

56 אותה קבוצה חלקית של אתרים היא קבועה לכל הפרופילים הגנטיים.

57 עבור Hash Table גדול, רוב השורות לא יהיו מאוכלסות מאחר ומספר הפרופילים הגנטיים האפשריים, גדול בהרבה ממספרם בפועל. גודל ה-Hash Table מוגבל על ידי כמות זיכרון המחשב.

3. התשובה לשאלה האם התוצאות המתקבלות במדגם תקפות ייחודית לאוכלוסיה ממנה נלקח המדגם, או בניסוח אחר, האם ניתן להכיל את התוצאות המתקבלות במדגם באופן ייחודי על האוכלוסיה ממנה נלקח המדגם, נובעת מאיכות המדגם עליו נעשית הבדיקה. כדי שניתן יהיה להכיל את תוצאות הבדיקה של המדגם באופן ייחודי על האוכלוסיה, המדגם צריך לענות על שני תנאים:

א. המדגם צריך לייצג באופן אמין את האוכלוסיה הנבדקת.

ב. המדגם צריך להיות שונה ממדגם אשר נלקח או יילקח מאוכלוסיה אחרת. קרי, לו היו נלקחים מדגמים, באותו האופן מאוכלוסיות שונות, אזי המדגמים אשר היו מתקבלים מהאוכלוסיות האחרות היו שונים מהמדגם שנלקח מהאוכלוסיה הנבדקת.

4. בבדיקת DNA לזיהוי אישי, האוכלוסיה היא הרצף המלא של מולקולת ה־DNA והמדגם הוא הפרופיל הגנטי.

5. בשיטה המקובלת לזיהוי אישי באמצעות פרופיל גנטי, ההתמודדות עם התנאי השני נעשה באמצעות הצגת השאלה ההסתברותית: מהו הסיכוי שיתקבל מאוכלוסיה אקראית מדגם זהה למדגם שהתקבל מהאוכלוסיה הנבדקת. היות והתוצאה המתקבלת היא בהסתברות של אחד למאות מיליונים או מיליארדים, כאשר מספר האוכלוסיות הקיימות בפועל במרחב הבדיקה לגביהן השאלה רלוונטית מסתכם בכמה מיליונים בודדים, אזי התשובה האינטואיטיבית היא כי סיכוי זה בטל בשישים, המדגם עונה על שני התנאים ולפיכך המדגם מייצג באופן אמין וייחודי את האוכלוסיה הנבדקת.

6. על בסיס תשובה אינטואיטיבית זו בתי המשפט קבעו את משקלה של ראיית ה־DNA, ראה ההפניה בפרק ההקדמה במאמר לדבריו של כבוד השופט חשין בפסק דין אבו חמאוד.

7. השאלה הראשונה הנבדקת במאמר המוצע היא: האם התשובה האינטואיטיבית שלעיל היא אכן נכונה ותקפה כאשר היא נבדקת באופן מושכל בכלים סטטיסטיים.
החישוב ההסתברותי מניב את התשובה המפתיעה כי התשובה האינטואיטיבית שגויה. מסקנת המאמר היא כי במספר האוכלוסיות הקיימות בפועל, במרחב הבדיקה לגביהן השאלה רלוונטית, קיימות אוכלוסיות נוספות אשר יכולות להניב את אותו המדגם בשיטה הדיגום המבוצעת. קרי, המדגם אינו עונה על התנאי השני לעיל.

8. זיהוי אישי באמצעות פרופיל גנטי מתבסס, בנוסף להנחת העבודה הגלויה, כי אין שני אנשים (למעט תאומים זהים) אשר יש להם מולקולת DNA זהה, גם על הנחת עבודה סמויה כי כאשר הבדיקה נעשית באמצעות מספר אתרים מספק, אין שני אנשים, במרחב הבדיקה, אשר יש להם פרופיל גנטי זהה.

9. המאמר מפריך את הנחת העבודה הסמויה ומראה באופן חד משמעי כי במרחב הבדיקה קיימים זוגות, שלישיות ורביעיות אשר יש להם פרופיל גנטי זהה במספרים משמעותיים גם כאשר הבדיקה נעשית באמצעות 9 אתרים.

10. יתרה מכך, המאמר מראה כי כאשר הבדיקה נעשית באמצעות 6 אתרים בלבד, התופעה של אנשים שונים בעלי פרופיל גנטי זהה קיימת במספרים עצומים.
לטעמי, מסקנות המאמר לגבי פרופילים גנטיים שהתקבלו באמצעות 6 אתרים הן כה מרחיקות לכת שיש מקום לבחון את תוקפן של ההרשעות שהתקבלו על בסיס בדיקות אלה.

11. מן ההיבט המעשי של תוצאות הבדיקה אני סבור שצריך להוסיף על ההיבטים התיאורטיים גם כמה היבטים מעשיים בכדי לבחון את ההשלכות התיאורטיות של המאמר על המסקנות המעשיות של משקלה של ראיית ה־DNA.
כפי שנקבע במאמר באמצעות נוסחה מספר 1, כמות הזוגות התואמים הממוצעת עולה בקירוב באופן ריבועי עם גידול האוכלוסיה.

$$\bar{n} = \frac{N \times (N-1)}{2} \times \bar{p}$$

לפיכך כדי לקבוע את כמות הזוגות התואמים, יש לקבוע את גודל האוכלוסיה N. הנחת המאמר היא כי גודל האוכלוסיה N הוא מספר תושבי מדינת ישראל כ־7 מיליון. אולם באופן מעשי הנחת עבודה זו אינה נכונה. לגבי כל העבירות לגביהן נשאלת השאלה האוכלוסיה N קטנה באופן משמעותי. אין מקום למחלוקת כי תינוקות, זקנים וחולים הרתוקים למיטתם, אסירים

בבתי סוהר, אזרחים הנמצאים בחו"ל במועד ביצוע העבירה וכיו"ב, אינם נכללים באוכלוסיה N. יתרה מכך, באותם המקרים בהם פרופיל ה-DNA מהזירה נקבע על תאי זרע, הגבול העליון של גודל האוכלוסיה N אלו רק זכרים אשר כבר/עדיין מייצרים תאי זרע.

12. אם בחקירת עבירה נתונה, גודל האוכלוסיה בפועל אליה יכול להשתייך מבצע העבירה היא בגודל של מיליון פרטים, אזי כמות הזוגות התואמים יורדת פי 49 מ-14,865 ל-303 זוגות תואמים בלבד.
לפיכך כדי לקבוע את משקלה של ראיית ה-DNA, לאור הממצאים במאמר, בכל מקרה נתון יש לאמוד את גודל האוכלוסיה האמיתי N, אליה יכול להשתייך מבצע העבירה, לקבוע את מספר הזוגות התואמים ולקבוע את משקל ראיית ה-DNA בהתאם.

13. הנחת העבודה בהשגה לעיל היא כי אין הבדל בערך הממוצע p בין האוכלוסיה הכללית לבין האוכלוסיה הספציפית N של העבירה הנדונה. באותם המקרים בהם מבצע העבירה משתייך לאוכלוסיה ייחודית עם ערך ממוצע של p שונה מהאוכלוסיה הכללית, ניתן לבצע את החישוב באמצעות ערכי N ו-p המתאימים ובאופן זה להתאים את מסקנות המחקר גם לאוכלוסיות ייחודיות (הערך p באוכלוסיות ייחודיות יהיה בדרך כלל גדול יותר מהערך p באוכלוסיה הכללית, או לפי הנחות המאמר — האוכלוסיה היהודית).
ראוי לציין בהקשר זה כי התאמת הערכים N ו-p לאוכלוסיות ייחודיות תשפיע באופן הפוך על n מספר הזוגות התואמים שיתקבלו. הקטנת N תקטין את n והגדלת p תגדיל את n.

14. עיון בחוות דעת מומחה מטעם התביעה מגלה כי במקרים רבים לא מתקבלות תוצאות חד משמעיות מכל האתרים הנבדקים. במקרים רבים כאשר הבדיקה נעשית באמצעות 9 אתרים מתקבלות תוצאות ב-8 או 7 אתרים בלבד והחישוב ההסתברותי של הערך p הספציפי, ההסתברות לקבלת הפרופיל הגנטי הספציפי באקראי, מתבסס על מספר האתרים בהם התקבלה התוצאה. כאשר זה המצב יש לחשב את הערך p הממוצע של האוכלוסיה רק על פי האתרים בהם התקבלה תוצאה בפועל. חישוב זה יגדיל את הערך הממוצע p ויגדיל את הערך n של מספר הזוגות התואמים. קביעת משקלה של ראיית ה-DNA צריך להיקבע בהתאם.

15. ראוי בהקשר של קבלת תוצאות רק בחלק מהאתרים הנבדקים, להאיר היבט נוסף, אשר ראוי למאמר עצמאי.
חישוב הערך p הספציפי במקרים של אונס, כאשר הדגימה מהזירה כוללת תערובת DNA מהקורבן והאנס, מבוצע בגישה המקובלת בישראל, על כל האללים המתקבלים וכוללת גם את האללים המשותפים לחשוד ולקורבן. גישה זו מתייחסת לפרופיל הגנטי של החשוד כהאפלוטייפ, לפיכך כוללים בחישוב גם את האללים המשותפים. כאשר התוצאה לא מתקבלת בכל האתרים מפעילים את אותה הגישה על האתרים בהם התקבלה תוצאה.
לדעתי יש כאן כשל קונספטואלי: על פי גישת ההאפלוטייפ לא ניתן לקבוע זהות בין הפרופיל הגנטי מהזירה לפרופיל הגנטי של החשוד כאשר לא מתקבלת תוצאה בכל האתרים הנבדקים, מאידך אם לא נוקטים בגישת ההאפלוטייפ אלא בגישת האתרים הבלתי תלויים, אזי לא צריך להכניס לתחשיב p הספציפי את האללים המשותפים לקורבן ולחשוד.
אם ננסה לשקלל את מסקנות המאמר ו/או ההערות לעיל, לקביעת משקלה של ראיית ה-DNA באותם המקרים של תערובת DNA עם תוצאות חסרות בחלק מהאתרים, יש להתייחס גם להיבט זה.

16. השאלה השנייה אליה מתייחס המאמר היא: האם האפשרות שאדם שהפרופיל הגנטי שלו מצוי במאגר נתונים גדול שבידי הרשויות, יופלל כתוצאה מפשע שביצע אדם שהפרופיל הגנטי שלו לא נמצא במאגר, היא אפשרות ממשית.
המאמר עונה על השאלה משני היבטים:

א. המאמר בודק באמצעות סימולציה את השאלה האם יש אפשרות ממשית שפרופיל גנטי של עבריין שאינו במאגר יימצא בזירת העבירה ויתאים לאדם, אשר לא ביצע את העבירה, אשר הפרופיל שלו נמצא במאגר.

ב. כמה פרופילים גנטיים צריכים להיות במאגר בכדי שיכלול פרופילים גנטיים זהים השייכים לאנשים שונים.

17. תוצאות החישוב של השאלה השנייה 57,413 ו-1,140 פרופילים הבנויים מ-9 ו-6 אתרים, בהתאמה, ממחיש באופן משכנע את הקושי המובנה בהתבססות על התוצאה המתקבלת מהתאמת הפרופיל מהזירה לפרופיל במאגר, הרי אם בפרופילים שנבדקו הנמצאים במאגר קיימים פרופילים זהים לאנשים שונים, אזי קל וחומר שקיימת אפשרות ממשית לקיומם של פרופילים זהים בין אנשים שנמצאים במאגר לעבריינים פוטנציאליים שאינם נמצאים במאגר.

18. להלן כמה הערות על הנחות העבודה בסימולציה שבוצעה בכדי לענות על שאלה א' לעיל.

הסימולציה מניחה את ההנחה הבאה: "מגרילים פרופיל גנטי אשר מדמה פרופיל גנטי של עבריין שאיננו נמצא במאגר..." על פי הנחת עבודה זו מתבצעת הסימולציה.

זוהי הנחת עבודה שגויה — היות וזהות העבריין אינה ידועה, צריך להניח כי קיימות שתי האפשרויות שהפרופיל של העבריין נמצא במאגר או שהפרופיל הגנטי של העבריין אינו נמצא במאגר. הנחת העבודה של הסימולציה שהפרופיל הגנטי של העבריין אינו נמצא במאגר, מאיינת את המשמעות של השאלה איזה משקל לייחס לעובדה שיש התאמה בין הפרופיל מהזירה לפרופיל במאגר.

19. בכדי לבחון את השאלה על פי הנחת העבודה שבסימולציה, קרי שהפרופיל הגנטי של העבריין אינו נמצא במאגר, צריך להכניס כמה שינויים בתהליך הסימולציה:

א. צריך לבנות קבוצה A הכוללת הפרטים במאגר (מיליון פרטים על פי המאמר) אשר להם פרופיל גנטי ידוע.

ב. צריך לבנות קבוצה B הכוללת את שאר הפרטים באוכלוסיה אשר הפרופיל הגנטי שלהם אינו ידוע. הפרטים בקבוצות A ו־B שונים ומהווים ביחד את כל האוכלוסיה.

ג. צריך להגריל לפרטים בקבוצה B פרופילים גנטיים באופן כזה שיילקחו בחשבון הפרופילים הגנטיים הידועים הנמצאים במאגר, כך שה־p הממוצע של צירוף שתי הקבוצות A ו־B יהיה ה־p הממוצע המיוחס לאוכלוסיה.

ד. להגריל באופן אקראי פרופיל גנטי מקבוצה B ולבחון אם הפרופיל קיים בקבוצה A.

ה. באמצעות מספר מספק של חזרות על תהליכי הסימולציה ניתן יהיה לקבוע קיימת אפשרות ממשית שהגרלת פרופיל גנטי מקבוצה B תניב פרופיל גנטי זהה לפרופיל קיים בקבוצה A.

20. למעשה זה היבט אחר של השאלה הראשונה במאמר או שאלת ימי ההולדת. השאלה היא מה הסיכוי שלפרט אקראי מקבוצה B יהיה תאריך הולדת זהה לאחד התאריכים המוגדרים בקבוצה A. היות והסיכוי לכך הוא גם פונקציה של גודלה של קבוצה B, גודל הקבוצה צריך להילקח בחשבון.

גישת הסימולציה: "מגרילים פרופיל גנטי אשר מדמה פרופיל גנטי של עבריין שאיננו נמצא במאגר" מגדירה את קבוצה B בגודל אינסופי (למעשה בגודל הערך ההופכי של p של קבוצה B), מה שאינו נכון במציאות.

21. גישה אפשרית חלופית היא לבנות את קבוצה B משתי תת קבוצות C ו־D, על פי הממצאים שהתקבלו בתשובה לשאלה הראשונה במאמר. קבוצה C כוללת את הפרטים עם הפרופילים הגנטיים שיש להם זוג תואם בקבוצה A וקבוצה D כוללת את קבוצת הפרטים עם הפרופילים הגנטיים שאין להם זוג תואם בקבוצה A (אני מתעלם מאפשרות של זוגות תואמים בתוך קבוצות C ו־D). בתהליך הסימולציה ניתן לאמוד את האפשרות להגריל באקראי מתוך קבוצה B פרופיל גנטי מקבוצה C. מה שמדמה את האפשרות למציאת פרופיל גנטי בזירה התואם פרופיל גנטי במאגר של פרט אשר לא ביצע את העבירה.

22. הצעת המאמר, לצמצום האפשרות למסקנות שגיאות מתוצאות בדיקות DNA, באמצעות הגדלת מספר האתרים הנבדקים, אכן יכולה לתת מענה הולם לבעיות המועלות במאמר.

ראוי להעיר בהקשר זה, כי לאור השיפורים הטכנולוגיים בקביעת רצף DNA, סביר מאוד כי בעתיד הנראה לעין ניתן יהיה לבצע זיהוי אישי באמצעות קביעת הרצף המלא של ה־DNA, בזמן ועלות סבירים, כך שהדיון בשאלת הזהות האקראית של פרופילים גנטיים לא יהיה רלוונטי.

ד"ר יורם פלוצקי (Ph.D. (Genetics, H.U.Jerusalem)., LL.B.(TAU).

הערת סיום – ד"ר מרדכי הלפרט

ברצוני להודות לסוקר על הערותיו למאמרי זה. מצאתי את הערותיו כמדגישות נקודות מעניינות ומצריכות הוספת הסברים בגוף המאמר עצמו. גרסת המאמר שפורסמה בסופו של דבר, כוללת בתוכה הסברים נוספים המתייחסים לנקודות שאותן העלה הסוקר.